

Modified Group Delay Based MultiPitch Estimation in Co-Channel Speech

Rajeev Rajan, Hema A. Murthy

*Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai, India
E-mail: { rajeev,hema } @cse.iitm.ac.in*

Abstract

Phase processing has been replaced by group delay processing for the extraction of source and system parameters from speech. Group delay functions are ill-behaved when the transfer function has zeros that are close to unit circle in the z -domain. The modified group delay function addresses this problem and has been successfully used for formant and monopitch estimation. In this paper, modified group delay functions are used for multipitch estimation in concurrent speech. The power spectrum of the speech is first flattened in order to annihilate the system characteristics, while retaining the source characteristics. Group delay analysis on this flattened spectrum picks the predominant pitch in the first pass and a comb filter is used to filter out the estimated pitch along with its harmonics. The residual spectrum is again analyzed for the next candidate pitch estimate in the second pass. The final pitch trajectories of the constituent speech utterances are formed using pitch grouping and post processing techniques. The performance of the proposed algorithm was evaluated on standard datasets using two metrics; pitch accuracy and standard deviation of fine pitch error. Our results show that the proposed algorithm is a promising pitch detection method in multipitch environment for real speech recordings.

Keywords: power spectrum, modified group delay, comb filter, spectrum estimation

1. Introduction

In speech and music research, robust pitch detection is a fundamental problem which finds many applications in day to day life. Pitch is the auditory attribute of a sound that allows its ordering on a frequency related scale. The rising and falling of pitch contours help in conveying prosody in speech and in tone languages, determine the meaning of words (Oxenham, 2012). A detail review on various monopitch estimation algorithms can be seen in (W.Hess, 1983; Rabiner et al., 1976; Gerhard, 2003). Pitch detection algorithms can be broadly classified into methods which operate in time domain, frequency domain, or both. The most commonly used time domain approaches are autocorrelation function and average magnitude difference function. In the frequency domain approaches, locating harmonic peaks is the key step in most of the algorithms (Schroeder, 1968). Studies show that in tonal languages, the relative pitch motion of an utterance contributes to the lexical information contained in a word unit (Gerhard, 2003). We cannot ignore the pitch information during recognition in such instances. A majority of the pitch tracking methods are usually

limited to clean speech and give a degraded performance in the presence of other speakers or noise. When a combination of speech utterances from two or more speakers are transmitted through a single channel, pitch cues of the individual sources will be weakened by the presence of mutual interference. In such ambiguous situations, estimating the accurate pitch tracks is a challenging task and currently is far from being completely solved, despite the attempts of several state-of-the-art approaches.

The multi-pitch estimation problem can be formulated as follows (Christensen et al., 2008): Consider a signal consisting of several, say K , sets of harmonics with fundamental frequencies ω_k , for $k = 1, \dots, K$, that is corrupted by an additive white Gaussian noise $\omega[n]$, having variance σ^2 , for $n = 0, \dots, N - 1$, i.e.,

$$x[n] = \sum_{k=1}^K \sum_{l=1}^L a_{k,l} e^{j\omega_k l n} + \omega[n] \quad (1)$$

where $a_{k,l} = A_{k,l} e^{j\phi_{k,l}}$ is the complex amplitude of the l^{th} harmonic of the source with $A_{k,l} > 0$, $\phi_{k,l}$ being the amplitude and the phase of the l^{th} harmonic of the k^{th} source respectively. The model in Equation (1) is known as the harmonic sinusoidal model. The task is to estimate the individual pitch estimates ω_k in the mixture signal. The estimation of the fundamental frequency, or the pitch of audio signals has a wide range of applications in Computational Auditory Scene Analysis (CASA), prosody analysis, source separation and speaker identification (de Cheveigne, 1993; Murthy and Yegnaranarayana, 2011). In music also, multipitch estimation is inevitable in applications such as the extraction of “predominant F_o ” (Salamon and Gomez, 2012), computation of bass line (Goto and Hayamizu, 1999), content-based indexing of audio databases (Tao Li et al., 2003) and automatic transcription (Ryynanen and Klapuri, 2008). Note that the interactive music applications demand highly robust real time pitch estimation algorithms in all aspects.

2. Related work

Numerous methods have been reported for multipitch estimation in speech and music (Li et al., 2008; Nishimoto et al., 2007; Wu and Wang, 2003). The correlogram based algorithm proposed by Wu *et al.* (Wu and Wang, 2003) uses a unitary model of pitch perception to estimate the pitch of multiple speakers. The input signal is decomposed into sub-bands using a gammatone filterbank and the framewise normalized autocorrelation function is computed for each channel. The peaks selected from all the channels are used to compute a likelihood of pitch periodicities and these likelihoods are modeled by a Hidden Markov Model (HMM) to generate the pitch trajectories. A subharmonic summation method and a spectral cancellation framework is used in the co-channel speech separation algorithm proposed by Li *et al.* (Li et al., 2008). Multi-pitch trajectory estimation based on harmonic Gaussian Mixture Model (GMM) and nonlinear Kalman filtering is also proposed for multipitch environments (Kameoka et al., 2004b). A constrained GMM based approach on the platform of information criterion is attempted in (Nishimoto et al., 2007).

In a polyphonic context, the overlap between the overtones of different notes and the unknown number of notes occurring simultaneously make the multipitch estimation a difficult and challenging task (Badeau et al., 2007). The algorithms used in polyphonic environment for pitch transcription include auditory scene analysis based methods (Kashino and Tanaka, 1993; Mellinger, 1991), signal model based Bayesian inference methods (Goto, 2004), unsupervised learning methods (Smaragdis and Brown, 2003; Virtanen, 2006) and auditory model based methods (Klapuri,

2008; Tolonen and Karjalainen, 2000; Wu and Wang, 2003). In auditory scene analysis based methods, acoustic features and musical information are used to group the sound sources present in a scene, while signal model based methods employ parametric signal models and statistical methods to transcribe the pitch tracks. Unsupervised learning techniques include independent component analysis, non-negative matrix factorization, usage of source-specific prior knowledge and sparse coding. In auditory model based methods, a peripheral hearing model is used for the intermediate data representation of the mixture signal, followed by periodicity analysis and iterative cancellation. Multi-pitch estimation in music can be used to extract various information such as number of simultaneous sounds, spectral envelopes and onset time/offset time of notes. If the pitch of a sound can be determined without getting confused by other co-occurring sounds, the pitch information can be used to organize simultaneous spectral components for their production (P.Klapuri, 2001).

Although the pitch is based on timing, it is hardly exploited for pitch estimation, primarily because phase appears to be noisy owing to the wrapping problem. On the other hand, group delay function that preserves the properties of the phase can be exploited. Group delay functions are poorly behaved when the signal is nonminimum phase. The modified group delay function was proposed in (Yegnanarayana and Murthy, 1992.) to address this issue. In this paper, this idea is extended to multipitch analysis. We propose a phase based signal processing algorithm as opposed to conventional magnitude based methods to retrieve the individual pitches in concurrent speech. The phase spectrum has to be first unwrapped before any meaningful analysis can be performed. The advantage of the group delay function instead of the phase spectrum is that it can be computed directly from the signal. Hence the problem of unwrapping of the phase spectrum can be solved.

The primary motivation for this work arises from the applications of the group delay function in estimating sinusoids from noise (Yegnanarayana and Murthy, 1992.). The algorithm starts from the flattened power spectrum. The modified power spectrum can be thought as a sum of sinusoids. This is then subjected to modified group delay processing to estimate the pitch components present in the speech mixture by iterative estimation and cancellation. Group delay based pitch extraction for a single voice is described in (Yegnanarayana et al., 1991).

The outline of the rest of paper is as follows. Section 3 explains group delay functions and modified group delay function briefly. The theory of pitch detection using modified group delay functions is described in Section 4. In Section 5, the proposed system for multi-pitch estimation is discussed in detail. Section 6 discusses the dataset and evaluation metrics followed by results and analysis in Section 7. The effectiveness of a variant of group delay feature is explained in Section 8. Conclusions are finally drawn in Section 9.

3. Group-delay functions and modified group delay functions (MODGD)

Signals can be represented in different domains such as time domain, frequency domain, z-domain and cepstral domain. In (Murthy, 1991), it was shown that signal information can be represented by group delay functions, one derived from the magnitude of the Fourier transform and the other from the Fourier transform phase.

Consider a discrete time signal $x[n]$. Then

$$X(e^{j\omega}) = |X(e^{j\omega})|e^{j\arg(X(e^{j\omega}))} \quad (2)$$

where $X(e^{j\omega})$ is the Fourier Transform (FT) of the signal $x[n]$ and $\arg(X(e^{j\omega}))$ is the phase function.

The group delay function $\tau(e^{j\omega})$ is defined as the negative derivative of the unwrapped Fourier transform phase with respect to the frequency.

$$\tau(e^{j\omega}) = -\frac{d\{\arg(X(e^{j\omega}))\}}{d\omega} \quad (3)$$

From Equation (2)

$$\arg(X(e^{j\omega})) = \text{Im}[\log X(e^{j\omega})] \quad (4)$$

Using Equation (3) and Equation (4), the group delay function can be computed directly from the signal as shown below (Oppenheim and Schaffer, 1990):

$$\tau(e^{j\omega}) = -\text{Im}\frac{d(\log(X(e^{j\omega})))}{d\omega} \quad (5)$$

$$\tau(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|X(e^{j\omega})|} \quad (6)$$

where the subscripts R and I denote the real and imaginary parts. $X(e^{j\omega})$ and $Y(e^{j\omega})$ are the Fourier transforms of $x[n]$ and $nx[n]$ respectively.

It is important to note that the denominator term $|X(e^{j\omega})|^2$ in Equation (6) becomes very small at zeros that are located close to the unit circle. This makes the group delay function very spiky in nature and also alters the dynamic range of the group delay spectrum. As the spikiness of the group delay function has no role to play in source/system characteristics, the computation of the group delay function is modified such that the source and system characteristics are not lost. The spiky nature of the group delay spectrum can be overcome by replacing the term $|X(e^{j\omega})|$ in the denominator of the group delay function with its cepstrally smoothed version, $S(e^{j\omega})$. The new function obtained is referred to as the modified group delay function in the literature. The algorithm for computation of the modified group delay function is described in (Hegde et al., 2007) and is given as

$$\tau_m(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|S(e^{j\omega})|^{2\gamma}} \quad (7)$$

where $S(e^{j\omega})$ is the cepstrally smoothed version of $X(e^{j\omega})$. The algorithm for the computation of MODGDF is given in (Murthy and Yegnanarayana, 2011). Two new parameters, α and γ are introduced to control the dynamic range of MODGDF such that $0 < \alpha \leq 1$ and $0 < \gamma \leq 1$. Modified group delay based algorithms can be used effectively to estimate system and source characteristics in speech processing (Murthy and Yegnanarayana, 2011).

4. Theory of pitch detection using modified group delay functions

The vocal tract system and its excitation contribute to the envelope and the fine structure respectively of the speech spectrum. The periodicity of the source manifests as picket fence harmonics in the power spectrum of the signal. If the vocal tract information can be suppressed, the picket fence harmonics are essentially pure sinusoids. The modified power spectrum can be thought as a sinusoidal signal. In the literature, it was shown that the modified group delay function is quite effective in estimating sinusoids in noise (Yegnanarayana and Murthy, 1992.). High resolution property of

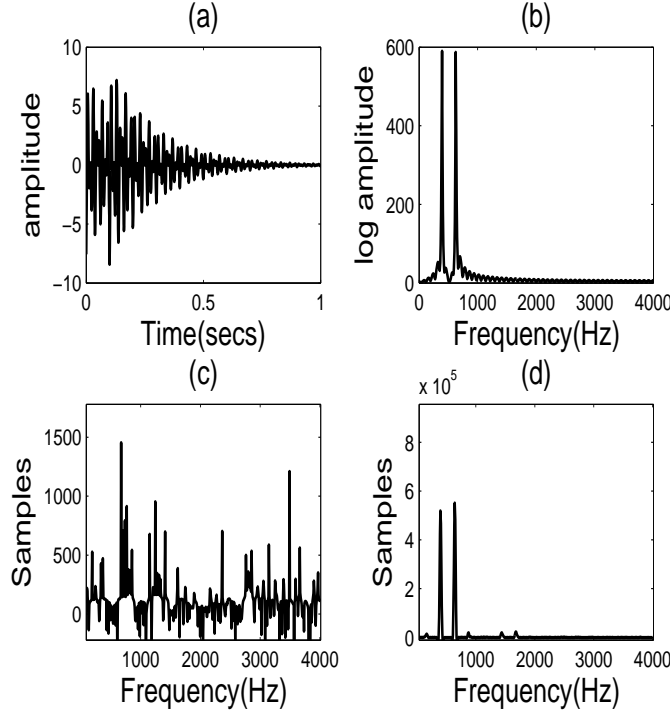


Figure 1: (a) Composite noisy signal, (b) Magnitude spectrum of a frame, (c) Group delay corresponds to frame in (b), (d) Modified group delay corresponds to frame in (b)

modified group delay (Murthy, 1991; Hegde, Rajesh M., 2005) is exploited in all those cases to resolve the spectral components. For instance, consider a noisy composite signal shown in Figure 1(a). Figure 1(b) shows its magnitude spectrum. Even though the group delay is spiky in nature (ref:-Figure 1(c)), spectral components are well resolved in the MODGD feature space in Figure 1(d). The monopitch estimation based on group delay function is explained in (Murthy, 1991). The process is illustrated in Figure 2 using the plots obtained in the intermediate steps. A frame of speech is shown in Figure 2(a). The flattened spectrum of the corresponding frame is shown in Figure 2(b). Peaks at multiples of fundamental frequencies can be observed in the MODGD plot shown in Figure 2(c). The peak in the MODGD feature space in the range corresponds to $[P_{min}, P_{max}]$ is mapped to the pitch estimate. The estimated pitch trajectory along with reference for an entire speech utterance is given in Figure 2(d). The systematic evaluation shows that the group delay based approach is at par with any other magnitude based approaches (Murthy, 1991).

The proposed method is an extension of the aforesaid process to multipitch environment. In the case of multiple speakers, the flattened power spectrum contains the excitation information of all the speakers. For instance, consider the z -transform of impulses separated by T_o and T_1 , corresponds to the excitation components, then

$$E(z) = 1 + z^{-T_o} + z^{-T_1} + z^{-2T_o} + z^{-2T_1} \quad (8)$$

The power spectrum of the source is given by

$$E(z)E^*(z) = (1 + z^{-T_o} + z^{-T_1} + z^{-2T_o} + z^{-2T_1})(1 + z^{T_o} + z^{T_1} + z^{2T_o} + z^{2T_1}) \quad (9)$$

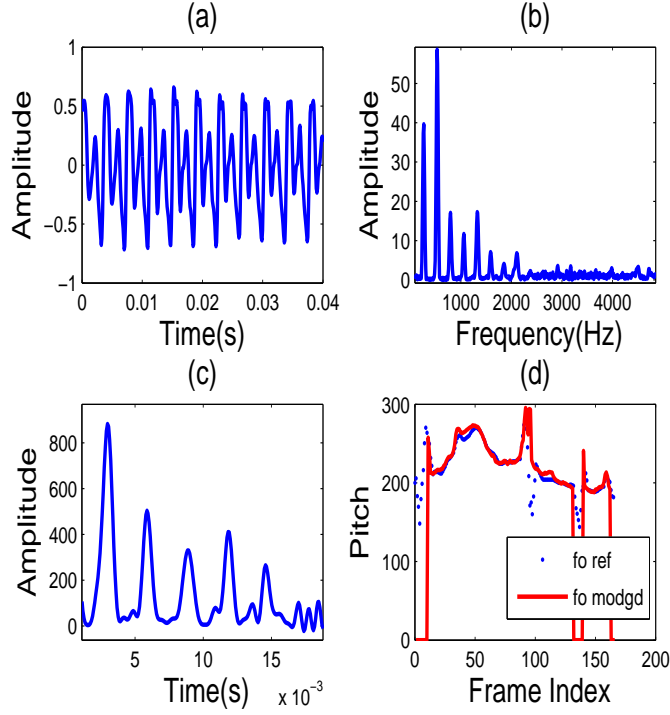


Figure 2: (a) Frame of a speech (b) Flattened power spectrum (c) Peaks in the MODGD feature space (d) Pitch estimated for the entire utterance with reference

Substituting $z = e^{j\omega}$,

$$\begin{aligned}
 |E(e^{j\omega})|^2 = & 5 + 4\cos(\omega T_o) + 4\cos(\omega T_1) + \\
 & 2\cos(\omega 2T_o) + 2\cos(\omega 2T_1) + 2\cos(\omega(T_o - 2T_1)) + \\
 & 2\cos(\omega(T_1 - 2T_o)) + 2\cos(\omega(2(T_1 - T_o))) + 2\cos(\omega(T_1 - T_o)) \quad (10)
 \end{aligned}$$

By restricting to three impulses per frame and evaluating the power spectrum on the unit circle as above and introducing a parameter γ , we have

$$|E(e^{j\omega})|^{2\gamma} = (3 + 2(1 + \cos(\omega T_o) + \cos(\omega T_1) + \cos(\omega(T_o - T_1)))^\gamma \quad (11)$$

where $0 < \gamma \leq 1$. The parameter γ controls the flatness of the spectrum. Thus the signal is a sum of sinusoids with frequencies that are integral multiples of $\frac{1}{T_o}, \frac{1}{T_1}, \frac{1}{T_o - T_1}$ and few combinations. If the spectral components corresponding to the periodic component are emphasised, the problem of pitch extraction reduces to that of the estimation of sinusoids in the frequency domain. We now replace ω by n and T_o, T_1 by ω_o, ω_1 in Equation (11) and remove the dc component to obtain a signal which is ideally a sum of sinusoids corresponds to excitation components present in the

mixture.

$$s[n] = \cos(n\omega_o) + \cos(n\omega_1) + \cos n(\omega_1 - \omega_o) + \cos(n2\omega_o) + \cos(n2\omega_1)..$$

$$n = 0, 1, 2, 3, \dots, N-1 \quad (12)$$

This signal is subjected to modified group delay processing, which results in peaks at multiples of the partials present in the speech mixture. The procedure to map this peak locations to constituent pitch trajectories is explained in the next section.

5. Proposed system description

The block diagram of the proposed system is shown in Figure 3. As seen in the Figure, the power spectrum of the speech signal is first flattened using cepstral smoothing technique to annihilate system characteristics by retaining the excitation information. In the mixed speech, the

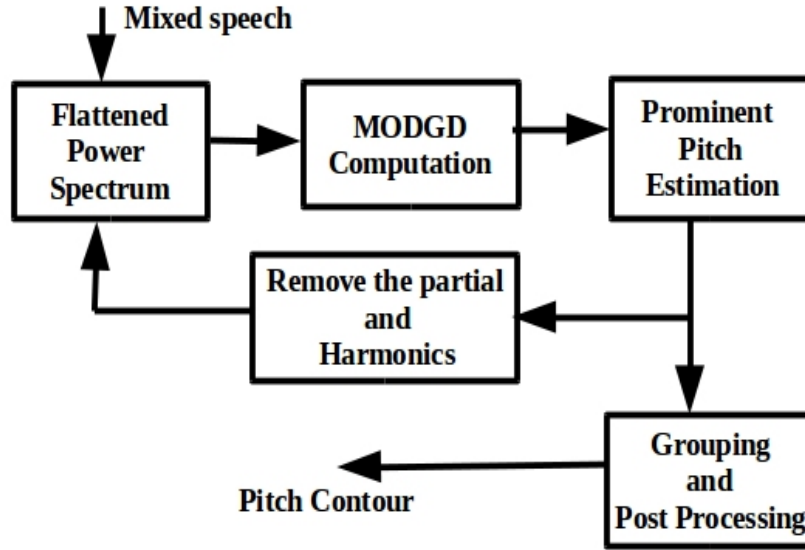


Figure 3: Block diagram of the proposed method

flattened spectrum consists of excitations of both the speakers. The flattened spectrum is frame-wise analysed using MODGD algorithm described in Section 3. As discussed in the Section 4, peaks can be seen in the MODGD feature space at locations corresponding to the multiples of all the pitch components and its few algebraic combinations. The location of the prominent peak in the range corresponds to $[P_{min}, P_{max}]$ in the MODGD feature space is mapped to the candidate pitch estimate in the first pass. In the second pass, the estimated pitch component and its harmonics are annihilated from the flattened power spectrum. Then the residual signal will be traced for the second frequency component using MODGD analysis. In the post processing phase, pitch grouping followed by removal of pitch outliers results in the final pitch trajectories. The subsequent sections describe these steps in detail.

5.1. FIR comb filtering

Once the prominent pitch is estimated from the flattened spectrum in the first pass, next we aim at the estimation of the second pitch candidate. In the second pass, the estimated pitch and its partials are removed from the flattened spectrum using a comb filter. Comb filters are widely used in many speech processing applications such as speech enhancement, pitch detection and speaker recognition (Jin et al., 2010), (Laskowski and Jinn, 2010). The FIR comb filter transfer function is given as :

$$H(z) = \frac{Y(z)}{X(z)} = 1 + \alpha z^{-D} \quad (13)$$

where D is the pitch period, α a constant and $X(z)$, $Y(z)$ represent the z -domain representation of input and output respectively. Magnitude response of the comb filter is

$$|H(e^{j\omega})| = \sqrt{(1 + \alpha^2) + 2\alpha \cos(\omega D)} \quad (14)$$

The basic structure of the comb filter and its responses are shown in Figure 4. In the proposed approach, comb filter is used to annihilate the predominant fundamental frequency component obtained in the first pass from a composite flattened power spectrum which constitutes multiple excitations.

For the instance, consider a speech mixture of two synthetic speech signals with f_0 s 200 Hz and 280 Hz. Modified group delay function computed for a synthetic mixture frame is shown in Figure 5. In Figure 5(a), blue color plot is the MODGD obtained in the first pass. The peaks in MODGD feature space, correspond to the pitch candidates present in the speech mixture and its integral multiples. In the first pass, the prominent peak in the MODGD feature space is mapped to first pitch estimate followed by the annihilation of it from the the residual spectrum. The red color contour in Figure 5(a) is the computed MODGD for the second pass. The individual pitch tracks computed through the aforesaid steps are shown in Figure 5(b) along with references. Similarly, another real audio mixture example is shown in Figure 6. Figure 6(a) shows the MODGD plot for a real audio frame and in Figure 6(b), pitch estimates of the audio segment are shown. The modified group delay functions obtained in the first pass and in the second pass are illustrated in the figure. It is obvious from the figure that the peak corresponds to the predominant pitch computed in the first pass is annihilated during the second pass.

5.2. Pitch trajectory estimation by grouping

At the end of the pitch estimation phase, two pitch candidates per frame are computed. In the pitch grouping stage, these candidates are grouped into trajectories which comprise continuous, smooth individual tracks. A more heuristic approach for grouping is the use of high-low criteria. Since pitch crossing is not considered, out of two candidates per frame high pitch values are grouped into one trajectory and low values to other.

Dynamic programming based pitch grouping can also be employed. In that case, the relative closeness of the distance between peaks in two consecutive frames is used to compute optimal path. Transition cost is computed as the absolute difference in distance between the current and previous frame. The optimal path is selected by minimizing the transition cost across frames using back tracking approach. The transition cost $C_t(c_j/c_{j-1})$ between the pitch candidates c_j and c_{j-1} of consecutive frames is given as (Veldhuis, 2000)

$$C_t(c_j/c_{j-1}) = |L_j - L_{j-1}| \quad (15)$$

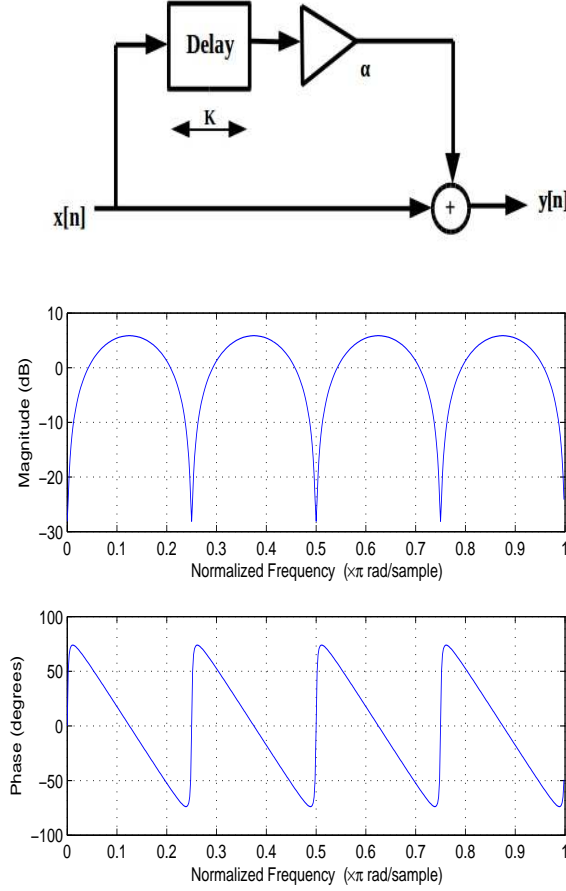


Figure 4: Comb filter structure with its magnitude and phase responses

where L_j, L_{j-1} are peak locations in consecutive frames. The dynamic programming algorithm finds an optimal pitch sequence $(c_1 \dots c_M)$ with candidates c_1 in the first and c_M in the M^{th} frame in a block by minimizing the transition cost function (Veldhuis, 2000). Transition cost $TC(c_1 \dots c_M)$ of pitch candidates c_1 to c_M is computed by

$$TC(c_1 \dots c_M) = \sum_{j=2}^{j=M} C_t(c_j/c_{j-1}) \quad (16)$$

The optimal sequence of pitch markers is determined by back tracking from the candidate c_M in the M^{th} frame in a block to its starting frame. If the pitch detection algorithm computes any spurious candidate, the dynamic programming may result in erroneous pitch tracks. The proposed algorithm is implemented using the first approach and form pitch contours by ensuring continuity.

5.3. Postprocessing

The accuracy in pitch estimation is improved by a post processing stage. In this stage, first task is to identify the segments where one or no speaker is present. A soft threshold on spectral flux is employed to identify these segments. The spectral flux is computed as the squared difference

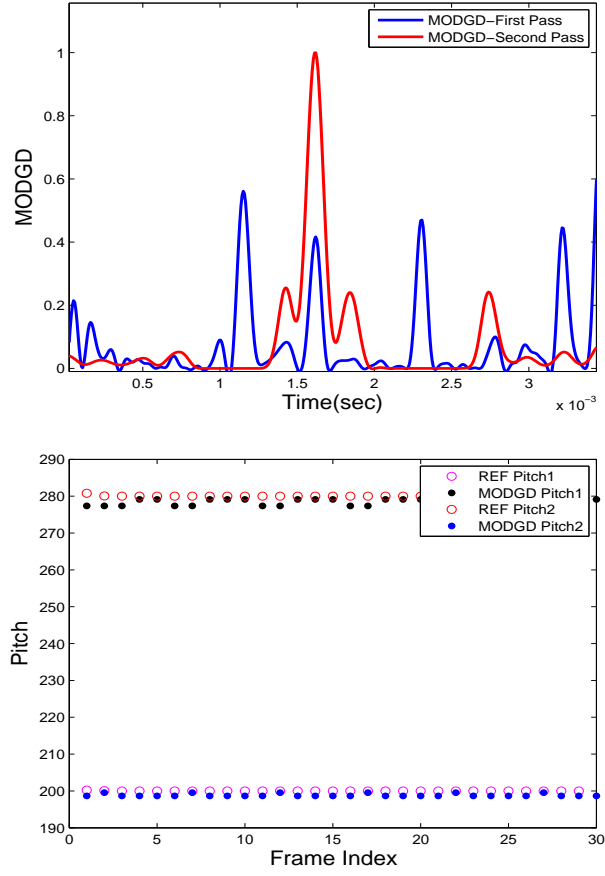


Figure 5: (a) MODGD on the Flattened spectrum for a frame (b) Pitch extracted for the mixed synthetic speech

between the normalized magnitudes of the spectral distributions of adjacent frames.

$$F_r = \sum_{k=1}^{N/2} (|X_r[k] - X_{r-1}[k]|)^2 \quad (17)$$

where $X_n[k]$ is the magnitude spectrum vector for the k^{th} subband of frame n . Segments which are detected as single speaker frames in the voicing detection stage is processed again for monopitch estimation using the MODGD algorithm. If the pitch estimated follows a path, estimated sequences are plugged into to the already formed contour by continuity check.

As part of smoothening the curve, stray values will be removed by framing rules to refine the pitch contour, thus minimizing the erroneous pitch estimates. For example, let f_t and f_{t+1} be the pitch candidates of consecutive frames in a pitch track after the grouping stage. If f_{t+1} lies outside the range $[f_t - \rho, f_t + \rho]$, this is treated as a spurious pitch estimate and will be interpolated using previous and successive pitch values (Radfar et al., 2011). We use linear interpolation for identifying missing pitch frequencies; however other interpolation techniques such as cubic or spline interpolation could be used. This simple but effective technique reduces the pitch error considerably. Note that missing pitch frequencies should typically not be interpolated for segments corresponding to 40 msec or longer for typical speech statistics (Radfar et al., 2011). The threshold ρ is set

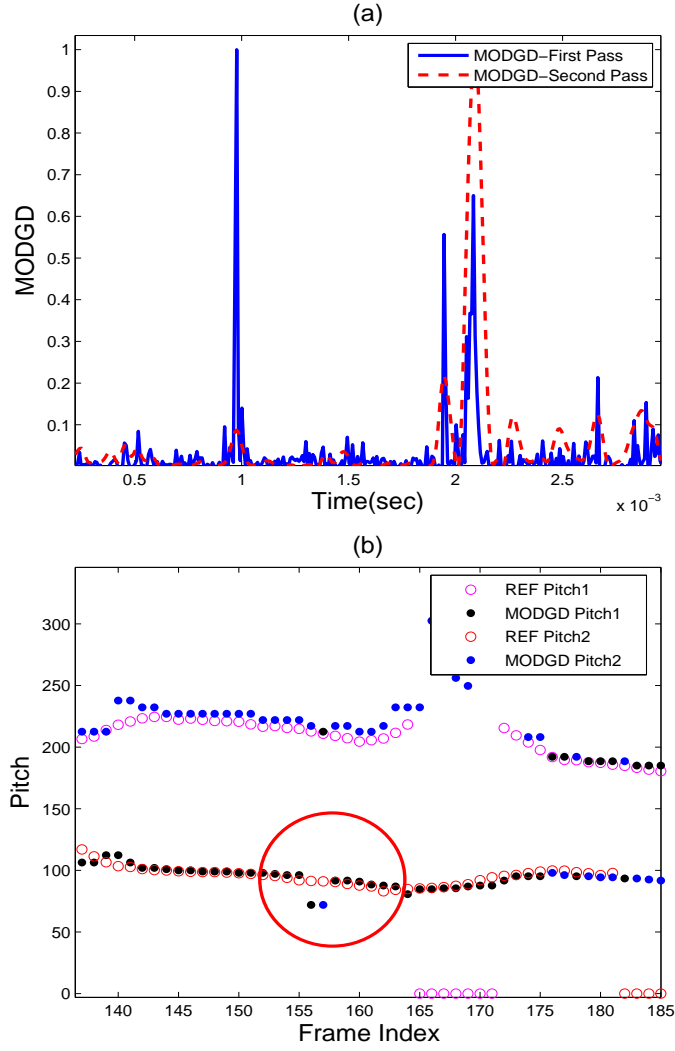


Figure 6: (a) MODGD plot for a frame in first pass (blue) and residual spectrum (red) for a real speech mixture, (b) Presence of stray values in the pitch contour are marked in circle

heuristically to 10 Hz. A typical example is shown in Figure 6(b). The circled part indicates the presence of two stray values in the middle of a continuous curve. The estimated pitch trajectories for a speech mixture with cross gender pattern is shown in Figure 7. Figure 7(a) shows the initial pitch estimates and Figure 7(b) shows the individual pitch trajectories after post processing. The pitch trajectories estimated using Wu *et al.* algorithm (D.L.Wang et al., 2003) is also shown in Figure 7(c) for the same speech mixture. A detail analysis of the results can be seen in the Section 7.

5.4. Pitch extraction in noisy and reverberant environment

The presence of noise and reverberation in speech poses major problems even in monopitch estimation. For noise corrupted speech, both the time-domain periodicity and spectral-domain periodicity are distorted and hence the conventional pitch estimation fails to certain extent (Huang and Lee, 2013). Group delay domain representation of speech makes it relatively immune to noise

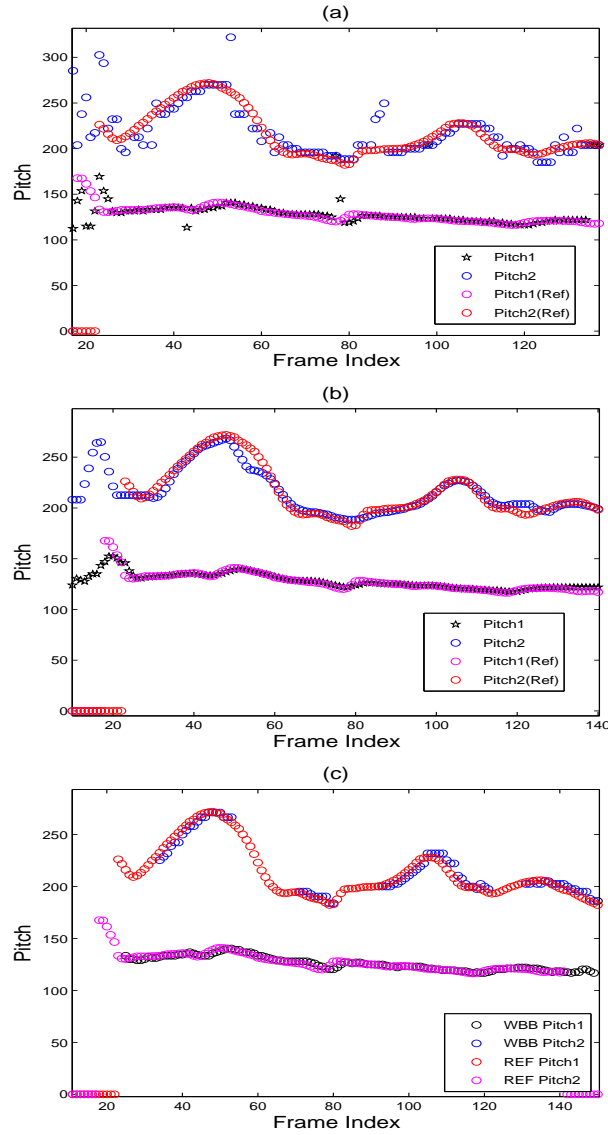


Figure 7: (a) Initial pitch estimates for a speech mixture (b) Final pitch trajectories estimated using the proposed algorithm (c) Pitch trajectories estimated using WWB algorithm

when compared to that of the short-time magnitude spectrum (Hegde et al., 2007; Yegnanarayana and Murthy, 1992.). For instance, consider the noisy signal $x[n]$ as the output of the autoregressive process $s[n]$, corrupted with Gaussian noise $\omega[n]$, i.e

$$x[n] = s[n] + \omega[n] \quad (18)$$

Group delay analysis of an autoregressive process in a noisy environment is given in (Yegnanarayana and Murthy, 1992.). Z transform of $s[n]$, ignoring the effects of truncation of the response of an all-pole system is given as

$$S(z) = \frac{GE(z)}{A(z)} \quad (19)$$

where $E(z)$ is the z transform of the excitation sequence $e[n]$ and $G/A(z)$ is the z transform of the all-pole system corresponding to the autoregressive process. From Equations (18) and (19)

$$X(z) = \frac{GE(z) + W(z)A(z)}{A(z)} = \frac{V(z)}{A(z)} \quad (20)$$

In group delay domain,

$$\tau_X(\omega) = \tau_V(\omega) - \tau_A(\omega) \quad (21)$$

As explained in (Yegnanarayana and Murthy, 1992.), the noise spikes in $\tau_X(\omega)$ can be suppressed by multiplying with the estimated zero spectrum. This results in an estimate of $-\tau_A(\omega)$ which corresponds to the spectral component in the composite signal. Thus group delay based approach is very effective in analyzing frequency components of a composite signal in the presence of noise. Room reverberation adversely affect the characteristics of pitch and thus makes the task of pitch determination more challenging. It causes degradation of the excitation signal due to the received speech signal because of the involvement of another filter which characterizes the room acoustics (Jin and Wang, 2010). In reverberant environments, the speech signal that reaches the microphone is superimposed with multiple reflected versions of the original speech signal. These superpositions can be modeled by the convolution of the room impulse response (RIR), that accounts for individual reflection delays, with the original speech signal (Allen and Berkley, 1979). Mathematically, the reverberant speech $r[n]$ is obtained as the convolution of speech signal $s[n]$ and room impulse response $h[n]$ (Thomas et al., 2008).

$$r[n] = s[n] * h[n] \quad (22)$$

Room impulse response is described as one realization of a non-stationary stochastic process in Schroeder's frequency-domain model (Jot et al., 1997) as

$$h[n] = b[n]e^{-\delta n}, \text{ for } n \geq 0 \quad (23)$$

where $b[n]$ is a centered stationary Gaussian noise, and δ is related to the reverberation time T_r . A typical room impulse response used for the experiment is shown in Figure 8. The proposed algorithm is also analysed in a reverberative condition using simulated impulse response.

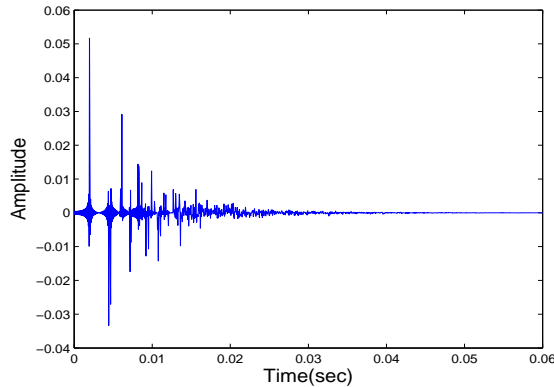


Figure 8: Room Impulse Response

Table 1: Category of mixtures for Dataset:1 and 2

Category	Speech data
1	Male/Female, Female/Female, Male/Male
2	Male/Female, babble noise
3	Male/Female, white noise
4	Male/Female with reverberation

6. Evaluation

6.1. Evaluation data set

In the proposed work, focus is given to the multipitch estimation of speech mixture with two speakers. The performance of the proposed algorithm was evaluated using following datasets:-

- Dataset-1: The dataset consists of 40 audio files obtained by mixing a subset of utterances from the Pitch Tracking Database of Graz university of Technology (PTDB-TUG) (Petrik et al., 2011), Childers database and a few audio samples from *Simple⁴All* speech corpora (Sun et al., 2014). The PTDB-TUG consists of audio recordings with phonetically rich sentences from TIMIT corpus. The TIMIT corpus consists of dialect sentences (labeled as *sa*), phonetically-compact sentences (labeled as *sx*), and phonetically-diverse sentences (labeled as *si*). *Simple⁴All* corpora consists of audio samples from different languages.
- Dataset-2: GRID (Cooke et al., 2006) is a large multitalker audiovisual sentence corpus to support joint computational-behavioral studies in speech perception. The corpus consists of high-quality audio and video recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). A subset of 40 audio files are used for generating mixtures for the evaluation.

In the experiments, each audio mixture is processed using a hamming window of frame length of 30 ms and hop size of 10 ms. As shown in Table 1, the interferences are classified into four categories by considering clean and noise conditions. The data set contains audio files of cross gender (male/female) and same gender (female/female, male/male) patterns. The test was conducted mainly on 0 dB target-to-masker ratio (TMR) which is considered the most difficult situation in co-channel speech segregation problem as both talkers equally mask each other. In category 2 and 3, speech is obtained by mixing the category 1 speech data with babble noise (5 dB SNR) and white noise (10 dB SNR). Category 4 interferences comprising of simulated reverberant speech utterances. The performance is also evaluated with speech mixture generated by clean voices of cross gender pattern with +3dB and -3dB Target to Masker Ratio(TMR).

Reverberant speech is generated using simulated room acoustics using a MATLAB implementation (Lehmann and Johansson, 2008) from the image model (Allen and Berkley, 1979). The model produces the room impulse response (RIR) when fed with room dimensions, wall reflection coefficients and physical locations corresponding to sound sources and the microphone. The simulation is done for reverberation time $T_{60} = 200\text{ms}$.

Table 2: Comparison of Accuracy (Dataset:1)

Category	$Accuracy_{20}(in\%)$			$Accuracy_{10}(in\%)$		
	MODGD	WWB	JIN	MODGD	WWB	JIN
Male-Female	88.52	77.95	81.99	84.58	76.71	81.04
Female-Female	85.28	64.54	72.00	75.02	60.78	70.31
Male-Male	80.53	66.24	72.41	73.58	66.01	70.01
Male-Female,Babble noise	84.68	60.56	76.90	72.12	60.17	73.43
Male-Female,White noise	78.04	63.11	77.59	73.13	62.97	76.21
Male-Female with reverberation	74.01	73.08	81.15	63.05	72.70	80.28

Table 3: Comparison of Accuracy (Dataset:2)

Category	$Accuracy_{20}(in\%)$			$Accuracy_{10}(in\%)$		
	MODGD	WWB	JIN	MODGD	WWB	JIN
Male-Female	87.88	79.58	79.95	82.65	78.92	78.95
Female-Female	78.99	77.17	77.30	74.86	76.74	76.81
Male-Male	74.39	50.92	73.50	65.76	50.84	73.08
Male-Female,Babble noise	77.40	57.29	74.09	70.00	56.25	72.74
Male-Female,White noise	65.68	66.96	72.66	58.73	66.34	71.64
Male-Female with reverberation	73.30	64.44	79.00	68.00	64.01	78.38

6.2. Evaluation metrics

The performance is evaluated only for voiced frames. The reference frequency of an unvoiced frame is considered as 0 Hz. To evaluate the performance of our algorithm, requires a reference pitch contour corresponding to the true individual pitch. We computed the reference pitch of clean speech using Wavesurfer (Ref, 2015). The guidelines for evaluating the performance of monopitch estimation can be seen in (Rabiner et al., 1976). Since there are no generally accepted guidelines for the performance evaluation in the case of multipitch tracking, we extended the guidelines of single pitch tracking. The performance is quantitatively assessed by measuring two types of metrics: accuracy and standard deviation of the fine pitch errors E_{fs} .

The metrics are defined as follows,

- **Accuracy:** $Accuracy_{10}$ and $Accuracy_{20}$ correspond to the percentage of frames at which pitch deviation is less than 10% and 20% with respect to the reference respectively. A gross error occurs if the detected pitch is not within the specified threshold with respect to the reference pitch.
- **Standard deviation of the fine pitch errors (E_{fs}):** The standard deviation of the fine pitch error is a measure of the accuracy of the pitch detection during voiced intervals. The standard deviation of the pitch detection σ_e is given as:

$$\sigma_e = \sqrt{\frac{1}{N} \sum (p_s - p'_s)^2 - e^2} \quad (24)$$

where p_s is the standard pitch, p'_s is the detected pitch, N is the number of correct pitch frames and e is the mean of the fine pitch error. e is given as:

$$e = \frac{1}{N} \sum (p_s - p'_s) \quad (25)$$

7. Analysis and Discussions

The performance of the proposed algorithm was evaluated primarily on speech mixtures, speaking simultaneously with equal average power. Three patterns, same gender (M/M, F/F) and cross gender (M/F) are considered for the evaluation. In addition to the clean speech condition, the performance is also evaluated in noisy and reverberant conditions. WWB algorithm (D.L.Wang et al., 2003) and Jin *et al.* algorithm (Jin and Wang, 2011) are used for the objective comparison in performance evaluation. The algorithm of Wu, Wang, and Brown is referred to as the WWB algorithm. WWB algorithm integrates a channel-peak selection method and Hidden Markov Model (HMM) for forming continuous pitch tracks. WWB framework computes final pitch estimates in three stages: auditory front-end-processing, pitch statistical modelling and HMM tracking. Jin *et al.* algorithm, designed specially to tackle reverberant noises is similar to WWB algorithm but different in channel selection and pitch scoring strategy. An auditory front-end and a new channel selection method are utilized to extract periodicity features in Jin *et al.* algorithm. In (D.L.Wang et al., 2003; Jin and Wang, 2011), half of the corpus is used to estimate the model parameters and thus supervisory in nature. Another important fact about the experiments reported in (D.L.Wang et al., 2003) is that they are focused on speech mixtures with one dominating speaker. Both this algorithms report considerable amount of transition errors, in which pitch estimates of speaker-1 are misclassified as the pitch estimates of speaker-2. For a fair comparison, the WWB and Jin *et al.* algorithm outputs are grouped in the post processing stage to ensure no transition error is occurred (Jin and Wang, 2011).

The grouping is done using a similar approach proposed in (Radfar et al., 2011). We consider each track as a cluster of data and the mean of each cluster as representative of that cluster. Let $g\{q\} = \{\omega^t, \dots, \omega^{(t+p)}, \dots, \omega^{(t+P-1)}\}$ be the q^{th} track (or equivalently cluster) with length P . Then the mean of the q^{th} cluster is defined as $M^q = \frac{1}{P}(\sum_{p=0}^{P-1} \omega^{(t+p)})$. For the two-speaker case, pitch tracks are classified into two groups (I and II), one belonging to each speaker. To do this, mean of the first segment in each track is computed as M^{1*} and M^{2*} respectively. Then successive segments are grouped into one of the tracks by assessing the closeness of M^q with M^{1*} and M^{2*} by fixing a threshold k .

The results obtained through quantitative evaluation are listed in Tables 2 - 7. Table 2 and 3 compare the pitch accuracies with 20% and 10% tolerance for dataset-1 and dataset-2 respectively. The results can be used to analyse the performance of the proposed system in clean and noisy conditions with same/cross gender speech mixtures. In clean conditions, the proposed group delay based system outperforms the other two systems. Jin *et al* algorithm and MODGD algorithm show a neck to neck performance giving slight advantage to MODGD system. Another important point we noticed in the experiment is that WWB algorithm fails to pick one of the pitch estimates in many frames. The proposed method reports accuracies of 84.58% and 82.65% within 10% tolerance for dataset 1 and dataset 2 respectively in clean mixtures. In noisy conditions, Jin *et al*, algorithm shows good performance especially in reverberant conditions. It is worth in noting that, in babble and white noise conditions, MODGD system is at par with the Jin *et al* algorithm and also shows a superior performance over WWB algorithm. In same gender mixture patterns, if the pitch values are too close, the performance of the proposed algorithm is affected due to the filtering operation. In the proposed group delay based system, both noise and source introduce zeroes that are close to the unit circle in the z domain (Murthy, 1991; Hegde, Rajesh M., 2005). The fundamental difference is that source zeroes are periodic while noise zeroes are aperiodic. This is the primary

reason why the proposed algorithm extracts pitch in the noisy environment. Even though in the anechoic condition, the proposed system and WWB algorithm yielding competitive performance, in reverberant environment, the performance of the proposed system is poor as compared to WWB algorithm. Table 4 and 5 compare the standard deviation of fine pitch error(E_{fs}) for dataset-1 and dataset-2 respectively. WWB algorithm and Jin *et al.* algorithms give E_{fs} in the range 2-4 Hz across the entire interference categories while the proposed algorithm reports slightly high E_{fs} in the range 3- 5.5Hz. Finally, we have done analysis on varying the Target to Masker Ratio (TMR) in the clean conditions. The results are tabulated in Table 6 and Table 7. The analysis shows a similar trend in the performance of both MODGD algorithm and Jin *et al* algorithm. A considerable variation in accuracy is reported in the case of WWB algorithm as TMR varies from -3dB to 3dB, but for the other two algorithms, the variation is not that much significant. When it comes to fine pitch E_{fs} , the variation over different TMR is minimal as compared to equal TMR situation.

Table 4: Comparison of E_{fs} corresponding to $Accuracy_{10}$ (in Hz) (Dataset:1)

Category	E_{fs}		
	MODGD	WWB	JIN
Male-Female	4.80	1.81	3.12
Female-Female	5.43	2.12	4.05
Male-Male	4.82	1.52	2.75
Male-Female,Babble noise	5.40	2.17	3.47
Male-Female, White noise	5.47	2.01	3.82
Male-Female with reverberation	4.98	2.87	3.76

Table 5: Comparison of E_{fs} corresponding to $Accuracy_{10}$ (in Hz) (Dataset:2)

Category	E_{fs}		
	MODGD	WWB	JIN
Male-Female	3.48	1.59	2.37
Female-Female	3.77	2.17	3.28
Male-Male	3.65	1.43	2.45
Male-Female, Babble noise	3.86	1.61	2.64
Male-Female, White noise	4.27	1.91	3.1
Male-Female with reverberation	4.34	2.55	3.00

Table 6: Comparison of accuracy in various TMR:Database:1

Category	WWB		JIN		MODGD	
	$Accuracy_{10}(in\%)$	E_{fs}	$Accuracy_{10}(in\%)$	E_{fs}	$Accuracy_{10}(in\%)$	E_{fs}
Male-Female 0dB	76.71	1.81	81.04	3.12	84.58	4.80
Male-Female -3dB	72.90	1.80	79.65	3.74	83.32	4.93
Male-Female +3dB	79.40	1.71	82.21	3.42	85.52	4.85

8. Source-MODGD cepstral features in estimating number of speakers.

In literature, many multipitch estimation algorithms start from the estimation of number of speakers. In (Kameoka et al., 2004b), a frame independent process is described, that gives good

Table 7: Comparison of accuracy in various TMR-Database:2

Category	WBB		JIN		MODGD	
	$Accuracy_{10}(in\%)$	E_{fs}	$Accuracy_{10}(in\%)$	E_{fs}	$Accuracy_{10}(in\%)$	E_{fs}
Male-Female 0dB	78.92	1.59	78.95	2.37	82.65	3.48
Male-Female -3dB	71.02	1.73	76.71	2.05	82.72	3.46
Male-Female +3dB	74.37	1.61	77.61	2.10	82.26	3.49

estimates of the number of speakers and f_{os} with a single-frame-processing. The algorithm explained in (Kameoka et al., 2004a) detects number of concurrent speakers based on maximum likelihood estimation of the model parameters using EM algorithm and information criterion. In the work proposed by S.Vishnubhotla *et al.* (Vishnubhotla and Espy-Wilson, 2008), the temporal evolution of the 2-D AMDF is used to estimate the number of speakers present in periodic regions. The proposed method can also be extended to speech mixture with more than two speaker, if the

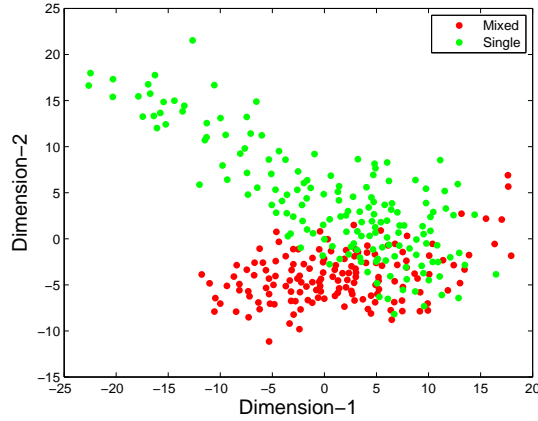


Figure 9: Two dimensional visualization of SMCC features for a single speaker and a speech mixture using Sammon mapping

information regarding the number of speakers is available. The iterative cancellation steps are determined by the number of speakers present in the mixture. Our experiments show that a variant of group delay feature; SMCC (Source-MODGD Cepstral features) derived from the flattened spectrum can be efficiently utilized to estimate the number of speakers. Since the modified group delay function behaves like a squared magnitude response (Murthy and Yegnanarayana, 2011), homomorphic processing approach can be employed to convert modified group delay spectra to meaningful features. The filter bank analysis on MODGD of the flattened power spectrum followed by DCT results in the proposed SMCC feature.

Steps to compute Source-MODGD Cepstral Coefficient features are summarized below

- Frame blocking the speech signal at a frame size of 20 ms and frame shift of 10 ms. A hamming window is applied on each frame.
- Speech power spectrum is flattened using the spectral envelop obtained by cepstral smoothing to annihilate the system characteristics.

Table 8: Confusion matrix for Multipitch Environment Task. (SP-1 denotes speech with single speaker, SP-2 denotes speech mixture with two speakers and so on). Class wise accuracy is given as the last column entry.

	SP-1	SP-2	SP-3	SP-4	%
SP-1	26	0	0	0	100
SP-2	2	20	3	1	77
SP-3	0	5	17	4	65
SP-4	0	3	8	15	58

- Apply MODGD algorithm on the flattened power spectrum to compute modified group delay function of the smoothed spectrum.
- Apply filter-bank on modified group delay $\tau_m(k)$ to get the Filter Bank Energies (FBEs).
- Compute DCT of log FBEs to get the SMCC feature vectors.

A multi- dimensional scaling technique, Sammon mapping (J. W. Sammon, 1969) is used to visualize the separability of SMCC features in Figure 9. Sammon mapping is a non-linear mapping of high dimensional feature vectors to low dimensional space based on gradient search. In the figure, SMCC features computed for a single speaker (SP-1) and a speech mixture (SP-3) are plotted. In the proposed method, 20 dimensional SMCC feature vectors are computed in the front-end using the steps described above. A Gaussian Mixture Model (GMM) based classifier is used in the classification stage. The feature vectors computed from the training set are used to build models for one-speaker case, two speakers case and so on. Out of 180 files available in the dataset, 60% files are used for training and the rest for testing. 12 component Gaussian mixture models (GMM) are used in the modelling different classes of the speech mixtures. During the testing phase, the classifier evaluates the likelihoods of the unknown speech mixture data against these models. The model that gives the maximum accumulated likelihood is declared as the correct match. The performance of the aforesaid feature was evaluated on speech mixtures generated by the subset of GRID dataset (Cooke et al., 2006). The results are tabulated as a confusion matrix in Table 8. The overall accuracy is 75 %. All the single speaker test utterances are classified correctly. The results show that the proposed feature is a promising one in estimating number of speakers in a mixed speech.

9. Conclusion

A phase based approach for multipitch estimation is presented in this paper, yielding competitive performance as compared to other state of the art approaches. In the proposed algorithm, the power spectrum is first flattened in order to annihilate the system characteristics. The flattened spectrum is processed using MODGD algorithm to estimate the predominant pitch in each frame in the first pass. Then the estimated pitch and its harmonics are filtered out using comb filter. In the second pass, the residual spectrum is again analysed using the group delay algorithm to estimate the second candidate pitch. The pitch grouping stage followed by the post processing step results in final pitch trajectories. The performance of the proposed algorithm was evaluated on speech mixtures with cross gender (Female, Male), same gender (Male/Male, Female/Female) patterns on

versatile datasets. The remarkable point in the proposed method is that the proposed method is an unsupervised approach using phase information. It does not require pre-training on source models from isolated recordings. The problem of estimation of number of speakers in a speech mixture is also addressed using a variant of group delay feature, Source-MODGD Cepstral Coefficient features and evaluated the performance using a subset of GRID corpus. The results obtained in the multipitch experiments show that the proposed algorithm is promising one in multipitch environment for real audio recordings.

10. Acknowledgement

The authors would like to thank M. Wu, W. Jin, D.L.Wang and Guy J. Brown for sharing their algorithm for multipitch estimation.

References

- Allen, J. B., Berkley, D. A., 1979. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer* 65 (4), 943–950.
- Badeau, R., Emiya, V., David, B., 2007. Multipitch estimation of quasiharmonic sounds in colored noise. in *Proc.of 10th Int. Conf. on Digital Audio Effects(DAFx)*.
- Christensen, M., Stoica, P., Jakobsson, A., Jensen, S. H., 2008. Multi-pitch estimation. *Signal Processing* 88 (4), 972–983.
- Cooke, M., Barker, J., Cunningham, S., Shao, X., May 2006. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Amer* 120, 2421–2424.
- de Cheveigne, A., 1993. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time domain cancellation model for auditory processing. *Journal of the Acoust. Soc. Am.* 93 (6).
- D.L.Wang, M.Wu, Brown, G., 2003. A multipitch tracking algorithm for noisy speech. *IEEE Trans. on Speech and Audio Signal Processing* 11, 229–241.
- Gerhard, D., 2003. Pitch extraction and fundamental frequency:history and current techniques, 0–22.
- Goto, M., 2004. A real-time music scene description system: predominant-Fo estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication* 43, 311–329.
- Goto, M., Hayamizu, S., May 1999. A real-time music scene description system: Detecting melody and bass lines in audio signals. *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, 31–40.
- Hegde, R. M., Murthy, H. A., Gadde, V. R. R., January 2007. Significance of the modified group delay features in speech recognition. *IEEE International Transactions on Audio,Speech and Language Processing* 15, 190–202.

- Hegde, Rajesh M., July 2005. Fourier transform based features for speech recognition. PhD dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Madras, India.
- Huang, F., Lee, T., January 2013. Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique 21 (1), 99–109.
- J. W. Sammon, J., March 1969. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computer C-18* (5), 401–409.
- Jin, W., Liu, X., Scordilis, M. S., Han, L., 2010. Speech enhancement using harmonic emphasis and adaptive comb filtering. *IEEE Trans. on Audio Speech and Language Processing* 18 (2), 356–368.
- Jin, Z., Wang, D., March 2010. A multipitch tracking algorithm for noisy and reverberant speech. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 4218–4221.
- Jin, Z., Wang, D., March 2011. HMM-based multipitch tracking for noisy and reverberant speech. *IEEE Trans. on Audio, Speech, Lang. Process* 19 (5), 1091–1102.
- Jot, J.-M., Cerveau, L., Warusfel, O., September 1997. Analysis and synthesis of room reverberation based on a statistical time-frequency model. In *103rd AES Convention*, New York.
- Kameoka, H., Nishimoto, T., Sagayama, S., 2004a. Accurate Fo detection algorithm for concurrent sounds based on EM algorithm and information criterion. In *Proceedings of Special Workshop in Maui (SWIM)*.
- Kameoka, H., Nishimoto, T., Sagayama, S., 2004b. Multipitch trajectory estimation of concurrent speech based on harmonic GMM and nonlinear kalman filtering. In *proceeding of INTERSPEECH 2004*, 2433–2466.
- Kashino, K., Tanaka, H., August 1993. A sound source separation system with the ability of automatic tone modeling. in *Proc. Int. Comput. Music Conf. International Computer Music Association*, 248–255.
- Klapuri, A., February 2008. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Trans. Audio Speech and Language Processing* 16 (2), 255–266.
- Laskowski, K., Jinn, Q., 2010. Modeling prosody for speaker recognition – why estimating pitch may be a red herring. *Proceedings of Odyssey 2010*.
- Lehmann, E., Johansson, A., July 2008. Prediction of energy decay in room impulse responses simulated with an image-source model. *J. Acoust. Soc. Amer* 124 (1), 269–277.
- Li, M., Cao, C., Wang, D., Lu, P., Fu, Q., Yan, Y., September 2008. Cochanel speech separation using multi-pitch estimation and model based voiced sequential grouping. In *proceeding of: INTERSPEECH 2008*.

- Mellinger, D. K., 1991. Event formation and separation of musical sound. Ph.D. dissertation, Stanford Univ., Stanford, CA.
- Murthy, H. A., December 1991. Algorithms for Processing Fourier Transform Phase of Signals. PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India.
- Murthy, H. A., Yegnanarayana, B., November 2011. Group delay functions and its application to speech processing. *Sadhana* 36 (5), 745–782.
- Nishimoto, T., Kameoka, H., Sagayama, S., 2007. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Trans. on Audio Speech and Language Processing* 15 (3), 982–994.
- Oppenheim, A. V., Schaffer, R. W., 1990. *Discrete Time Signal Processing*. Prentice Hall, Inc, New Jersey.
- Oxenham, A. J., September 2012. Pitch perception. *The Journal of Neuroscience* 32 (39), 13335–13338.
- Petrik, M. W. S., Pirker, G., Pernkopf, F., 2011. A pitch tracking corpus with evaluation on multipitch tracking scenario. in *Proc. Interspeech*, 1509–1512.
- P. Klapuri, A., May 2001. Multipitch estimation and source separation by the spectral smoothness principle. *Acoustics, Speech and Signal Processing (ICASSP)*, 2001 IEEE International Conference on 5, 3381–3384.
- Rabiner, L., M.J.Cheng, A.E.Rosenberg, A.McGonegal, 1976. A comparative study of several pitch estimation algorithms ASSP-23.
- Radfar, M., Dansereau, R., Chan, W., Wong, W., May 2011. Mptracker: A new multi-pitch detection and separation algorithm for mixed speech signals. *Acoustics, Speech and Signal Processing (ICASSP)*, 2011 IEEE International Conference on, 4468 –4471.
- Ref, 2015. <http://www.speech.kth.se/wavesurfer/>. Wavesurfer-URL.
- Ryynanen, M., Klapuri, A., 2008. Automatic transcription of melody ,base line,and chords in polyphonic music. *Computer Music Journal* 32 (3), 72–86.
- Salamon, J., Gomez, E., August 2012. Melody extraction from polyphonic music signals using pitch contours characteristics. In *IEEE Transactions on Audio Speech and Language Processing* 20 (6), 1759–1770.
- Schroeder, M. R., 1968. Period histogram and product spectrum: New methods for fundamental-frequency measurement 43 (4), 829–834.
- Smaragdis, P., Brown, J. C., October 2003. Non-negative matrix factorization for polyphonic music transcription. in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 177–180.

- Suni, A., Raitio, T., Gowda, D., Karhila, R., Gibson, M., Watts., O., September 2014. The simple4all entry to the blizzard challenge 2014. In Proc. of the Blizzard Challenge 2014 Workshop.
- Tao Li, A., Ogihara, M., Li, Q., 2003. A comparative study on content-based music genre classification. In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 1759–1770.
- Thomas, S., Ganapathy, S., Hermansky, H., 2008. Recognition of reverberant speech using frequency domain linear prediction. *IEEE Signal Processing Letters* 15, 681–684.
- Tolonen, T., Karjalainen, M., 2000. A computationally efficient multipitch analysis model. *IEEE Trans.Speech and Audio Processing* 8 (6), 708–716.
- Veldhuis, R., October 2000. Consistant pitch marking. Proc. Sixth International Conf.on Spoken Language Processing 3, 207–210.
- Virtanen, T., 2006. Unsupervised learning methods for source separation in monaural music signal. in *Signal Processing Methods for Music Transcription*, 267–296.
- Vishnubhotla, S., Espy-Wilson, C., 2008. An algorithm for multi-pitch tracking in co-channel speech. *INTERSPEECH*.
- W.Hess, 1983. Pitch determination of speech signals: Algorithms and devices.
- Wu, M., Wang, D., 2003. A multipitch tracking algorithm for noisy speech. *IEEE Trans. Speech Audio Processing* 11 (3), 229–241.
- Yegnanarayana, B., Murthy, H. A., 1992. Significance of group delay functions in spectrum estimation. *IEEE Trans. Signal Process*, 40, 2281–2289.
- Yegnanarayana, B., Murthy, H. A., V.R.Ramachandran, May 1991. Processing of noisy speech using modified group delay functions. In Proc. of the IEEE Int. Conf. on Audio, Speech and Signal Processing, 945–948.